Rare Heart Transplant Rejection Classification Using Diffusion-Based Synthetic Image Augmentation

Han Bao[†], Jie Deng[†], Shihao Xing[†], Yishan Zhong[†], Wenqi Shi[†], Benoit Marteau[†], Bibhuti Das[‡], Bahig Shehata[§], Shriprasad Deshpande[¶], and May D. Wang[†]
[†] Georgia Institute of Technology, Atlanta, GA 30332
email: maywang@gatech.edu
[‡] University of Mississippi Medical Center, Jackson, MS 39216
[§] Wayne State University School of Medicine, Detroit, MI 48201
[¶] Children's National Health System, Washington, DC, 20010

***Abstract*—Heart Transplant Rejection (HTR) is a rare condition that requires early detection to prevent lasting damage to the transplanted heart. Unfortunately, the current HTR grading through biopsy image classification lacks consistency among pathologists. In addition, it is a time-consuming task. In this work, we have developed an automated diagnosis pipeline to streamline the heart transplant histopathology image quantification and classification, in order to provide objectivity for clinical decision support for pathologists. Traditionally, developing an automated image classification requires a substantial amount of labeled data. However, HTR is a rare condition and the dataset is usually unbalanced. For example, the dataset from DNA Based Transplant Rejection (DTRT) comprises 1,509 rejection tile images and 190 times more non-rejection tile images. To address the small data sample challenge in training the classifiers, we developed a novel strategy that used diffusion model to generate synthetic images of rejection. We conducted comprehensive HTR grade classification comparing results using dataset with synthetic rejection tiles versus the dataset without any synthetic rejection tiles. The introduction of synthetic augmentation resulted in an improvement from 0.781 to 0.981 for sensitivity, and an improvement from 0.984 to over 0.998 in AUROC. This study illustrated that synthetic data augmentation is a feasible strategy in developing AI solutions for rare diseases. In the future, we will expand in this direction to benefit more rare disease clinical decision support development.**

***Index Terms*—histopathology, whole slide image, diffusion, augmentation, generative model**

I. Introduction

A. Background and Motivation

Pediatric heart transplant rejection (HTR) is a lethal condition requiring prompt identification and response [1]. Thus, to prevent long-term damage to the transplanted heart, efficient and accurate diagnosis is critical. Currently, clinicians rely on invasive and risky biopsies of the heart tissue, stained with Hematoxylin and Eosin (H&E). The whole slide images (WSIs) are examined under a microscope to identify cellular features indicating early signs of rejection. This diagnostic process is burdensome for patients, requiring multiple biopsies, and carries the risk of missing subtle signs of rejection due to subjective interpretations of biopsy results. Moreover, there is low agreement among experts when analyzing the same set of biopsies [2].

To address these challenges and improve patient and clinician experience, researchers have turned to artificial intelligence (AI) for decision support. Over the years, several automatic systems utilizing deep learning (DL) models have been developed for WSI clinical problems, including heart transplant rejection detection [3]–[9]. These models are trained using images generated through digital pathology. However, a significant issue in training these models arises from the imbalance between rejection and non-rejection training data. In addition, the rarity of the disease limits the data availability. Both factors make the model prone to overfitting to the majority class (non-rejection), hamper learning from the minority class (rejection), compromise model generalization, and mislead evaluation metrics [10].

Therefore, it is critical to address the scarcity and imbalance of training images for heart transplant rejection detection in DL classification models. We propose augmenting the limited rejection training images using synthetic image generation techniques, such as Generative Adversarial Networks (GANs) [11] and Denoising Diffusion Probabilistic Models (DDPMs) [12], also simply known as diffusion models. Our hypothesis is that robust data augmentation with synthetic images for classes with few instances will benefit the deep learning-based rejection detection process. We compared the classification performance under three scenarios: 1) baseline binary classification without synthetic data. 2) with synthetic rejection tiles generated with diffusion model. 3) with synthetic rejection tiles generated with progressive GAN (PGAN) [13]. Our work is outlined in Figure 1.

II. Method

A. Data

Our dataset contains whole slide images obtained from various patients. These images were acquired from 40x magnified H&E stained and paraffin-fixed pediatric endomyocardial biopsies from two hospitals: DNA Based Transplant Rejection (DTRT) and Children's Healthcare of Atlanta (CHOA). For simplicity, in subsequent discussion, we refer to the two hospitals as Hospital A and Hospital B, respectively. The acute cellular rejection (ACR) region within each slide was manually

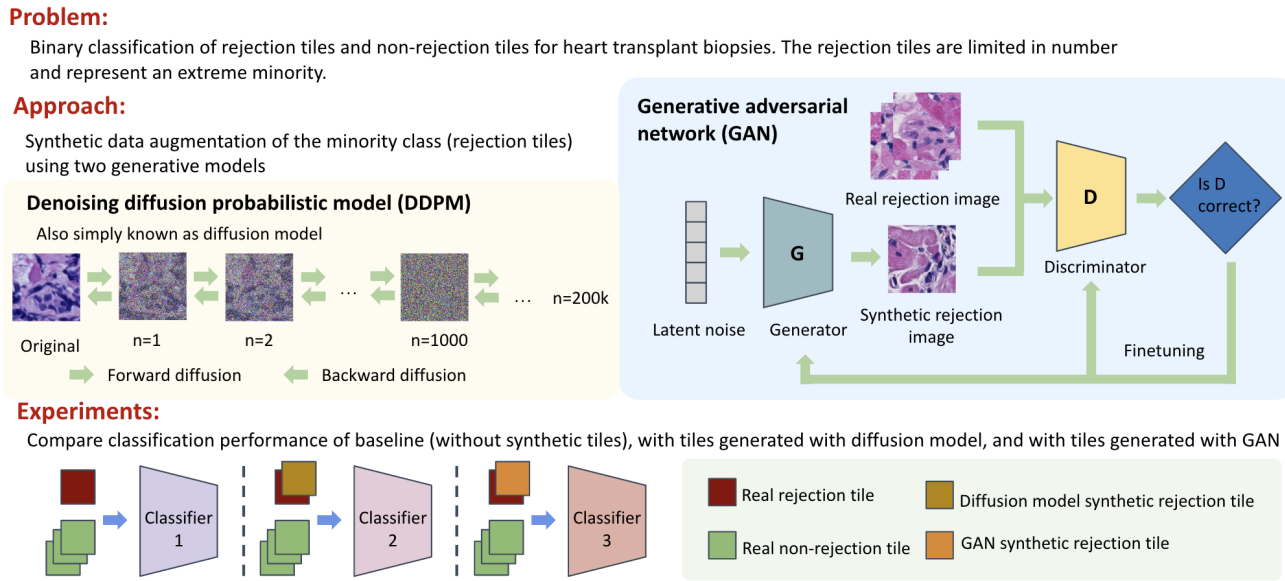

Fig. 1: Overview of the pipeline. We compared the impacts of incorporating synthetic rejection tiles in classifying rejection and non-rejection tiles for heart transplant histopathology.

labeled by two experienced doctors. The dataset consists of a total of 143 WSIs.

B. Pipeline and Methodology

First, we preprocessed our slide-level images to obtain tile-level images. Next, we employed transfer learning by utilizing these tiles as our baseline dataset to fine-tune a pre-trained EfficientNet-b3 classification model, which yielded baseline classification results. To address the challenge posed by the severely limited rejection tiles, we applied the two generative models on rejection tiles. Afterward, we repeated the fine-tuning procedure on the augmented data and compare the classification results with those of the baseline. Finally, we evaluated the performance of our methods both qualitatively and quantitatively.

C. Preprocessing

Each WSI was cut into 256 $\times$ 256 patches with a 50% overlap between neighboring patches in both the vertical and horizontal axes. We then calculated the ACR overlap ratio, and rejection tiles were then selected within rejection slides with ACR overlap $\geq$ 0.6, and non-rejection tiles were selected within non-rejection slides with ACR overlap = 0. After the above pre-processing steps, we have 287,106 non-rejection tiles and 1,509 rejection tiles from Hospital A; 12,167 non-rejection tiles, and 105 rejection tiles from Hospital B.

To avoid data leaking in classification training, we carefully selected our training, testing, and validation set. At slide level, we used WSIs from Hospital **A** as our training and validation set and WSIs from Hospital **B** as testing. The training-validation split ratio was 75:25.

Given that our testing set and training/validation set originated from different hospitals, it was necessary to address the difference in the percentage of rejection tiles between the two sets. In the baseline set, we randomly excluded a portion of non-rejection tiles, while in the augmented set, we introduced additional generated rejection tiles to balance the ratio and compensate for the limited representation of rejection tiles. This approach helped maintain a comparable distribution of rejection tiles across training, validation, and testing sets for both baseline and augmented datasets. The distribution for tile-level images is shown in table I.

Source	Labels	# of Tiles	
		Baseline	Augmented
Hospital A	Rejection	1,509	2,478
	Non-Rejection	174,857	287,106
Hospital B	Rejection	105	105
	Non-Rejection	12,167	12,167

TABLE I: Tile-Level Distribution: Baseline & Augmented Datasets

D. Synthetic Rejection Tiles Generation

The below generative models are selected and trained with rejection tiles from Hospital A.

1) Diffusion Model: Diffusion models [12] are a class of likelihood-based generative models that work by first adding noise to an image (forward) and then gradually removing the noise (backward), step by step. The backward process is trained to remove noise from images, with the objective expressed as a weighted variational bound. To generate a new synthetic image with a trained model, random noises are added to an image consecutively, and the trained model recovers the original image step by step, a process very similar to model training. Because the noises added are random, each full cycle of forward and backward passes would result in a different generated image. Specifically for our diffusion model, we experimented with different diffusion iterations (steps) to create images of 128 $\times$ 128. We then resized and doubled the widths and heights. Creating 256 $\times$ 256 images directly would have led to significantly longer training time.

2) PGAN: Progressive GAN [13], like the original GAN [11], consists of a generator and a discriminator. However, PGAN training starts with generating low resolution synthetic images and progressively generate synthetic images with high and higher resolution. Specifically, we started with the dimension of 4×4 after resizing the from the original tile image size of 256×256, and the width and height doubled until reaching back to 256×256.

E. Choice of Classification Network

For our particular task dealing with a large number of images, we decided to fine-tune the popular and efficient EfficientNet-b3 network [14] as our classification network. EfficientNet-b3 offers the best trade-off in terms of accuracy and model size. Our EfficientNet-b3 was initialized with Image-Net parameters, and we replaced the original 1000-dimensional layer with a 2-dimensional layer for our binary classification task. We used a batch size of 24, Adam optimizer, and a learning rate of 1×10^{-5}.

III. Results

A. Evaluation Metrics

We referred to [15] for our choices of evaluation metrics. To evaluate the performance of our augmentation methods, we used Fréchet inception distance (FID) score [16] and Spatial Fréchet Inception Distance (sFID) score [17] as our metric. The FID score is a metric that calculates the distance between feature vectors calculated for real and generated images in the Inception-V3 [18] latent space. The sFID score is another version of the FID score but uses spatial features rather than standard pooled features. In general, the lower the FID score or the sFID, the better the quality of generated images. We use FID as our main metric as it captures both diversity and fidelity, and we include sFID score to further validate the FID score.

B. Augmentation Results & Analysis

To quantitatively evaluate the capability of the diffusion and the PGAN model, we calculated the FID and sFID scores of the generated rejection tiles against the baseline rejection tiles. For the diffusion model, we collected 2000 rejection tiles generated from models generated after 100,000, 150,000, and 200,000 iterations. For the PGAN, we collected 2000 rejection tiles from the model after the full progressive training process. The results are shown below in Table II

	FID	sFID
PGAN	85.472	81.071
Diffusion (100,000 iterations)	21.226	70.319
Diffusion (150,000 iterations)	16.800	69.835
Diffusion (200,000 iterations)	14.362	69.349

TABLE II: FID Scores and sFID Scores for Generated Images

Typically, to obtain reliable FID scores, it is preferable to have a reference set with more than 10,000 sample images. Having a larger sample size improves the accuracy of FID scores, whereas insufficient sample images typically lead to lower FID scores. Furthermore, it should be noted that the Inception-V3 model was trained specifically on the ImageNet dataset, which might not capture the precise features of medical images. Despite these limitations, we can conclude that the diffusion model outperformed the PGAN model based on the significantly lower FID score and sFID score achieved by the images generated by the diffusion models. Additionally, the observed monotonic decrease and subsequent plateauing in FID and sFID as training iterations increased for the diffusion model shows that the model converged at 200,000 iterations.

A visual comparison between the real tiles and generated tiles from both methods is shown below in Fig. 2.

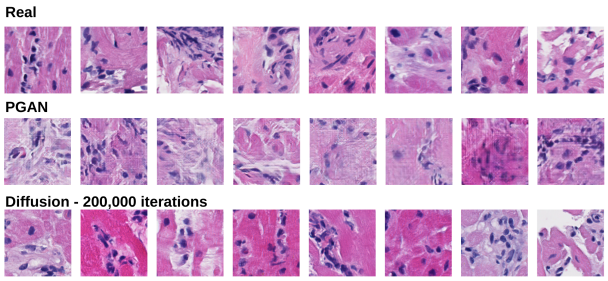

Fig. 2: Qualitative overview of the real rejection tiles, GAN-synthesized tiles, and diffusion-synthesized tiles. Visually, it is evident that the images generated by PGAN have a slightly lower quality.

C. Classification Results & Analysis

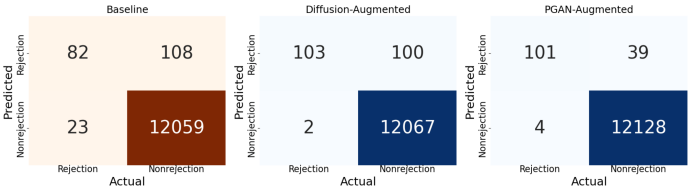

Fig. 3: Confusion Matrices for models trained from baseline, Diffusion, and PGAN-augmented datasets

	Baseline	Diffusion-augmented	PGAN-augmented
Sensitivity	0.781	0.981	0.962
Specificity	0.991	0.992	0.997
Precision	0.432	0.507	0.721
Accuracy	0.989	0.992	0.997
F1 score	0.556	0.669	0.825
MCC	0.576	0.702	0.832
AUROC	0.984	0.997	0.998

TABLE III: Metrics Comparison for models trained from baseline, Diffusion, and PGAN-augmented datasets

Sensitivity is important as it represents the true positive rate, indicating the ability to accurately detect rejection regions without overlooking any rejection region - missing rejection regions can have serious implications in terms of diagnostic accuracy and patient care. The diffusion-augmented dataset achieves a sensitivity of 0.981, outperforming both the original dataset (0.781) and the PGAN augmented dataset (0.962). While augmenting the training dataset with either diffusion or PGAN-generated images leads to better results, this indicates that the diffusion-augmented approach beats the PGAN-augmented approach in correctly identifying rejection regions,

which is essential for our task as we aim to minimize the chances of overlooking potential rejection areas.

Furthermore, AUROC provides an overall measure of the classification performance, revealing the model's ability to discriminate between rejection and non-rejection instances across different decision thresholds. The diffusion and PGAN-augmented methods yield an AUROC value of 0.997 and 0.998 respectively, surpassing the original dataset (0.984) and demonstrating that models from the augmented approach, either by diffusion or PGAN-generated images, show superior discriminative power and classification accuracy, outperforming the baseline model.

Also, the effect of augmentation can be compared through explainable AI techniques. Figure 4 presents a sample of 10 correctly predicted rejection images and the GradCAM++ Visualization for classification models trained with diffusion and PGAN-augmented dataset. Notably, the GradCAM++ heatmaps clearly indicate that the classification model trained with diffusion-augmented data can identify regions of interest over the nuclei with higher accuracy and precision when compared to the model trained with PGAN-augmented data, highlighting the enhanced sensitivity achieved by the diffusion-based augmentation approach in capturing relevant features for rejection classification.

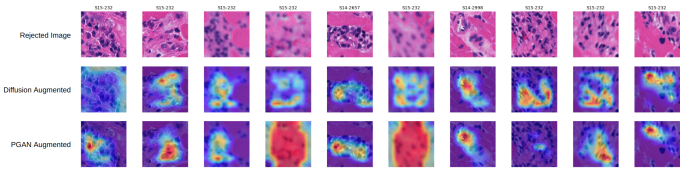

Fig. 4: Correctly predicted rejection images and GradCAM++ visualization

IV. Conclusion

We present a novel approach to identifying rare pediatric heart rejections using WSIs. Due to the rarity of the condition, we have an extremely imbalanced dataset consisting of 1,614 rejection tiles and 299,273 non-rejection tiles. Despite the unusual dataset, we achieved a robust classification and proved that synthetic image augmentation improves the classification performance from 0.781 sensitivity to 0.981 sensitivity and 0.984 AUROC to over 0.998 AUROC. Our findings further suggest that augmenting with diffusion-generated image yields superior sensitivity compared to PGAN. In the era of big models and big data, the scarcity of data shall not be a hindrance for rare diseases. Synthetic image augmentation using the state-of-the-art diffusion model is shown to be a promising solution. We genuinely hope that our work could inspire future applications of diffusion models for other rare diseases.

V. Acknowledgements

This work was supported by grants from Enduring Hearts, Petit Faculty Fellow Award, Wallace H Coulter Fellow Award, and Amazon Research Award to Professor Wang.

References

[1] M. E. Richmond, S. D. Zangwill, S. J. Kindel, S. R. Deshpande, J. N. Schroder, D. P. Bichell, K. R. Knecht, W. T. Mahle, M. A. Wigger, N. A. Gaglianello *et al.*, "Donor fraction cell-free dna and rejection in adult and pediatric heart transplantation," *The Journal of Heart and Lung Transplantation*, vol. 39, no. 5, pp. 454–463, 2020.

[2] H.-M. Yang, C. K. Lai, D. W. Gjertson, T. Baruch-Oren, S. H. Ra, W. Watts, W. D. Wallace, P. Shintaku, J. A. Kobashigawa, and M. C. Fishbein, "Has the 2004 revision of the international society of heart and lung transplantation grading system improved the reproducibility of the diagnosis and grading of cardiac transplant rejection?" *Cardiovascular Pathology*, vol. 18, no. 4, pp. 198–204, 2009.

[3] F. O. Giuste, R. Sequeira, V. Keerthipati, P. Lais, A. Mirzazadeh, A. Mohseni, Y. Zhu, W. Shi, B. Marteau, Y. Zhong *et al.*, "Explainable synthetic image generation to improve risk assessment of rare pediatric heart transplant rejection," *Journal of Biomedical Informatics*, vol. 139, p. 104303, 2023.

[4] A. Mirzazadeh, A. Mohseni, S. Ibrahim, F. O. Giuste, Y. Zhu, B. M. Shehata, S. R. Deshpande, and M. D. Wang, "Improving heart transplant rejection classification training using progressive generative adversarial networks," in *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 2021, pp. 1–4.

[5] J. Ye, Y. Luo, C. Zhu, F. Liu, and Y. Zhang, "Breast cancer image classification on wsi with spatial correlations," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 1219–1223.

[6] R. Feng, X. Liu, J. Chen, D. Z. Chen, H. Gao, and J. Wu, "A deep learning approach for colonoscopy pathology wsi analysis: accurate segmentation and classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 10, pp. 3700–3708, 2020.

[7] T. P. Seraphin, M. Luedde, C. Roderburg, M. van Treeck, P. Scheider, R. D. Buelow, P. Boor, S. H. Loosen, Z. Provaznik, D. Mendelsohn *et al.*, "Prediction of heart transplant rejection from routine pathology slides with self-supervised deep learning," *European Heart Journal-Digital Health*, vol. 4, no. 3, pp. 265–274, 2023.

[8] L. Tong, R. Hoffman, S. R. Deshpande, and M. D. Wang, "Predicting heart rejection using histopathological whole-slide imaging and deep neural network with dropout," in *2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, 2017, pp. 1–4.

[9] Y. Zhu, L. Tong, S. R. Deshpande, and M. D. Wang, "Improved prediction on heart transplant rejection using convolutional autoencoder and multiple instance learning on whole-slide imaging," in *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, 2019, pp. 1–4.

[10] P. Shamsolmoali, M. Zareapoor, L. Shen, A. H. Sadka, and J. Yang, "Imbalanced data learning by minority class augmentation using capsule adversarial networks," *Neurocomputing*, vol. 459, pp. 481–493, 2021.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[12] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[13] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[14] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," 2020.

[15] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," 2021.

[16] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," 2018.

[17] C. Nash, J. Menick, S. Dieleman, and P. W. Battaglia, "Generating images with sparse representations," 2021.

[18] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.